

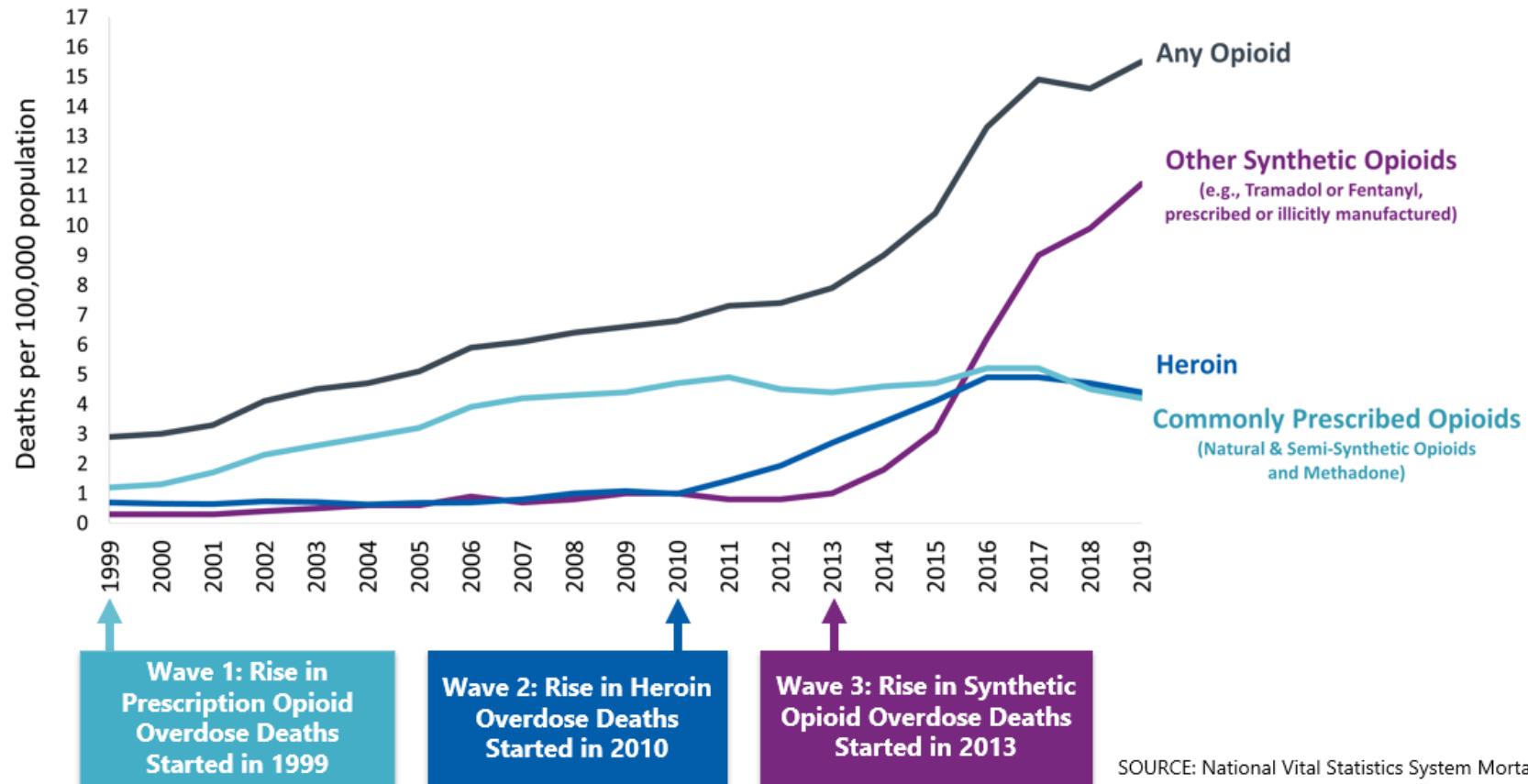
Original explainability Slides

AI Ethics: Explainability of Machine Learning

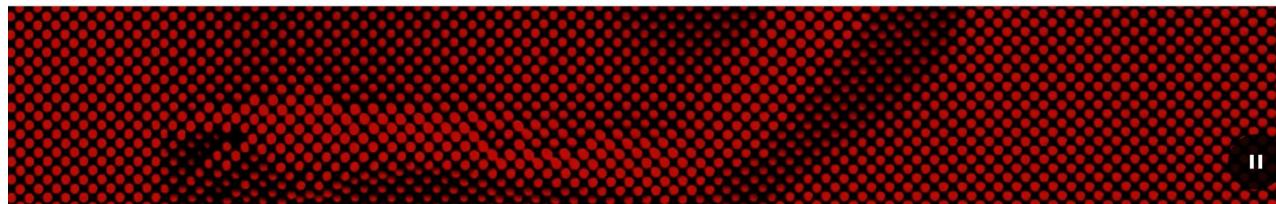
CS229: Machine Learning
Carlos Guestrin
Stanford University

©2022 Carlos Guestrin

Three Waves of the Rise in Opioid Overdose Deaths



SOURCE: National Vital Statistics System Mortality File.



VIDEO: SAM CANNON

MAIA SZALAVITZ

BACKCHANNEL AUG 11, 2021 6:00 AM

The Pain Was Unbearable. So Why Did Doctors Turn Her Away?

A sweeping drug addiction risk algorithm has become central to how the US handles the opioid crisis. It may only be making the crisis worse.

[The AI Database →](#)

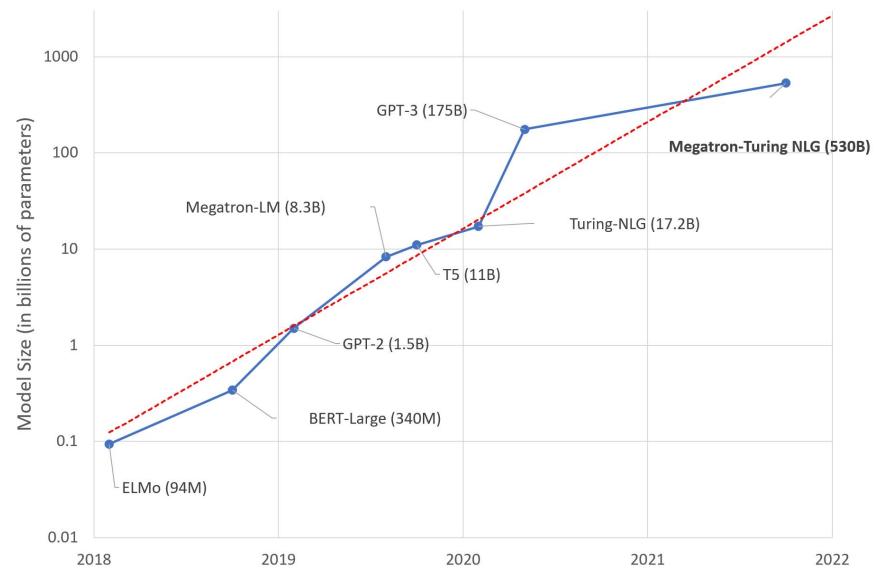
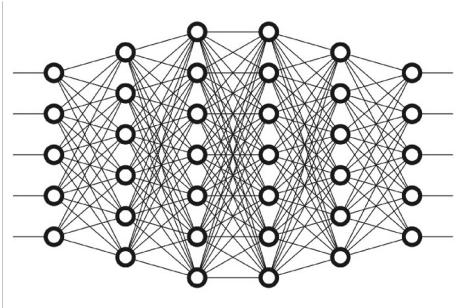
APPLICATION: ETHICS, PREDICTION, REGULATION

SECTOR: HEALTH CARE, PUBLIC SAFETY

ONE EVENING IN July of 2020, a woman named Kathryn went to the hospital in excruciating pain.

A 32-year-old psychology grad student in Michigan, Kathryn lived with endometriosis, an agonizing condition that causes uterine-like cells to abnormally develop in the wrong

ML Models More and More Complex



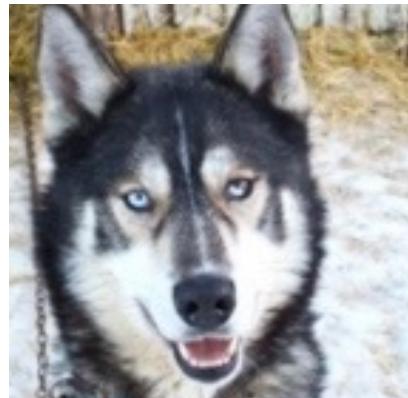
When is a model ready to deploy?

*Hard to understand when models are working
(for the right reasons) and not working!!*

Isn't test accuracy enough?

A User Study on Test Accuracy

Train a neural network to predict **wolf** v. **husky**



Husky



Wolf

VIDEO

SLATE IN MOTION.

OCT. 14 2016 3:18 PM

The Man Who Accidentally Adopted a Wolf Pup

It did not go well.

By A.J. McCarthy



Train a neural network to predict **wolf** v. **husky**

					
Predicted: wolf True: wolf	Predicted: husky True: husky	Predicted: wolf True: wolf	Predicted: wolf True: husky	Predicted: husky True: husky	Predicted: wolf True: wolf

Explanations for neural network prediction



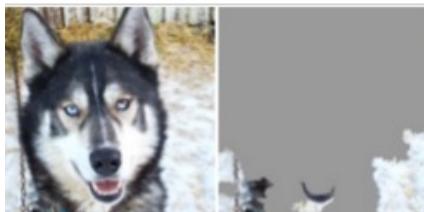
Predicted: **wolf**
True: **wolf**



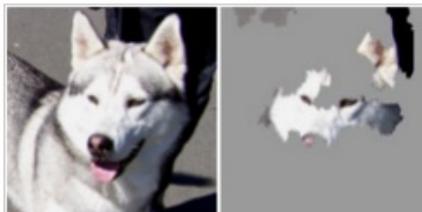
Predicted: **husky**
True: **husky**



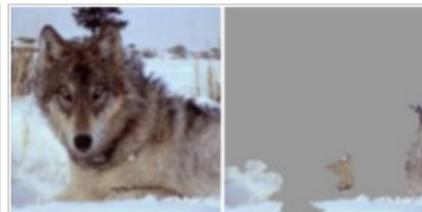
Predicted: **wolf**
True: **wolf**



Predicted: **wolf**
True: **husky**



Predicted: **husky**
True: **husky**



Predicted: **wolf**
True: **wolf**

Test accuracy may not capture critical issues

- Bad data
- Biases
- Poor performance in critical cases
- ...

Examining Models

Debugging is One Reason to Examine Models

- Examining models:
 - Why a model makes particular predictions
 - What alternative predictions are possible
 - How robust/stable are predictions
 - What data supports predictions
- Examining models for debugging: discover bad, unexpected or unstable behavior
 - Typically not discovered by accuracy in train/test data

Examining Models to Detect Algorithmic Bias

- Evaluate multiple fairness criteria
- Verify how/if decisions depend on sensitive features
- Discover what groups are privileged/disadvantaged by predictions

Examine Models for Recourse

- In opioid overdose risk case, patient deemed risky had no way to discover why
 - Or how to fix bad data
- Understanding why could enable individuals to:
 - Address data issues
 - Change their actions to change outcomes

Score Summary as of 11/18/2020

[Print this Report & Score](#)

Where You Stand



The Equifax Credit Score™ ranges from 280-850. Higher scores are viewed more favorably.

Your 3 credit scores are calculated by Equifax using the information contained in your Equifax, Experian, and TransUnion credit reports.

Equifax & Experian & TransUnion: Your score is considered **excellent**. Based on this score, you should be able to qualify for some of the lowest interest rates available and a wide variety of competitive credit offers should be available to you.

Learn more in [Understanding Your Score](#)

EFX EXP TU
▼
760 - 850
Excellent
37%

Range	280 - 559	560 - 659	660 - 724	725 - 759
Poor		Fair	Good	Very Good
US Population	12%	21%	18%	12%

What's Impacting Your Scores

Below are the key areas from these credit reports that are impacting your scores. [About Credit Scores](#)



Interpretable Models vs Post-hoc Explanations

Interpretability in ML

*Giving humans a **mental model** of
the machine's model behavior*

Learning Interpretable Models

(c.f., Lethan & Rudin 2015)

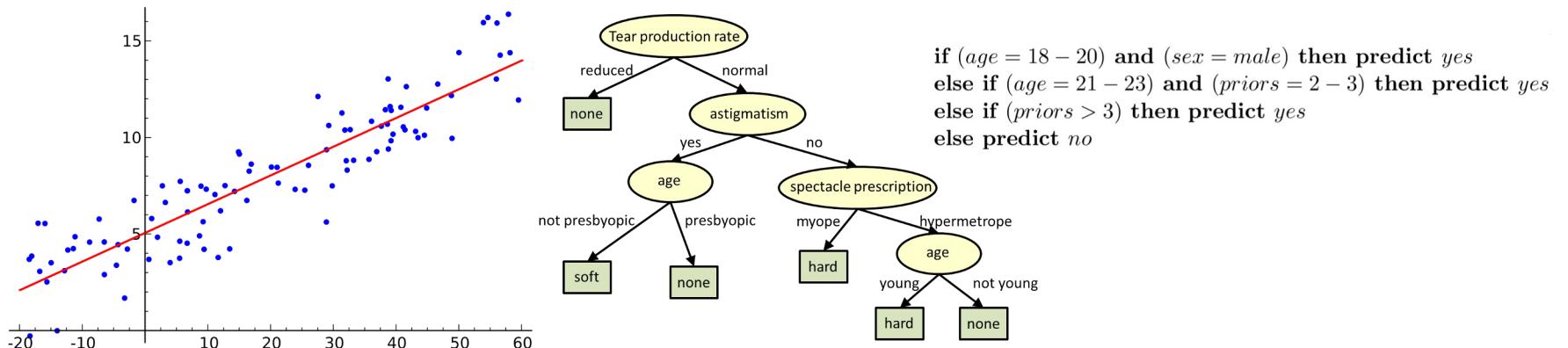
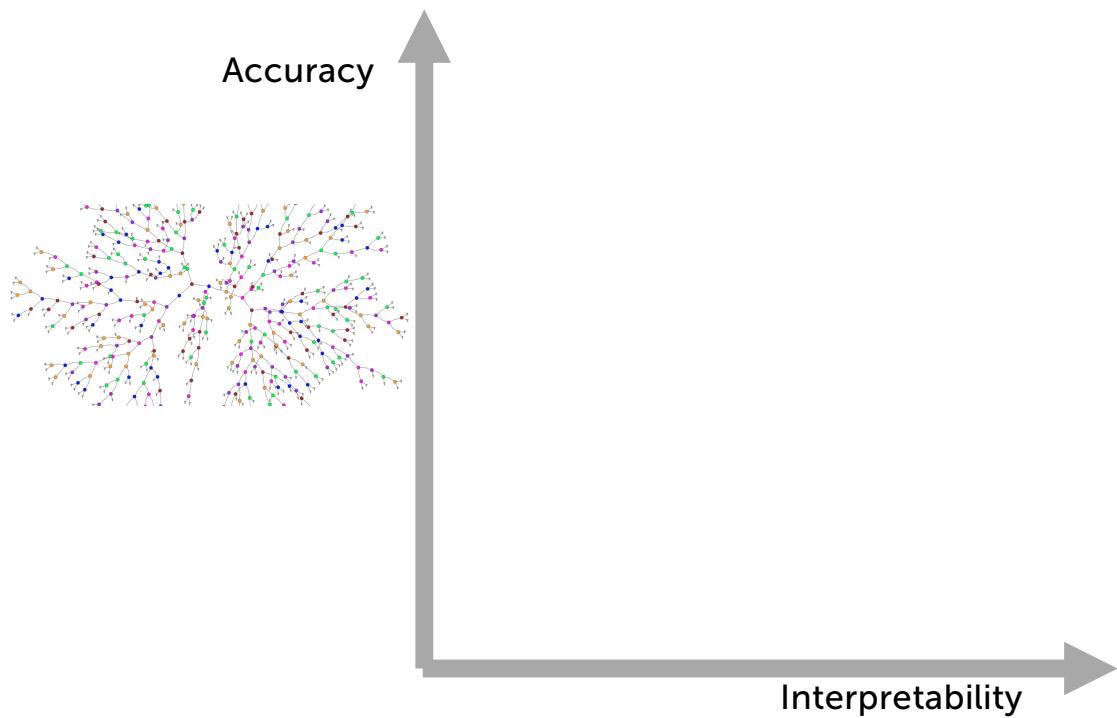
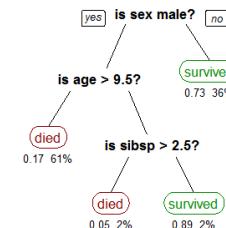


Image credit: Lakkaraju, Adebayo, Singh NeurIPS 2020 Tutorial

Accuracy vs Interpretability



©2022 Carlos Guestrin



CS229: Machine Learning

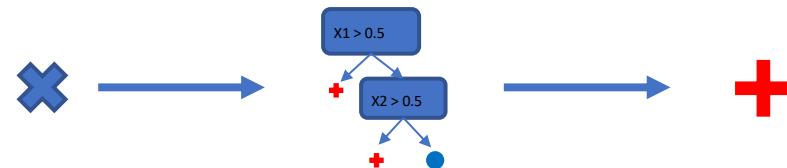
Post-hoc Explanations

- Given a (huge, complex) model, provide human explanations for predictions

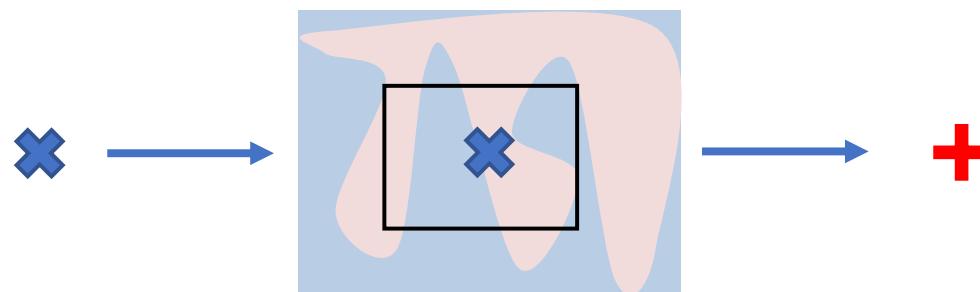


LIME: Local, Interpretable Model-Agnostic Explanations

Model agnostic → Ignore any internal structure

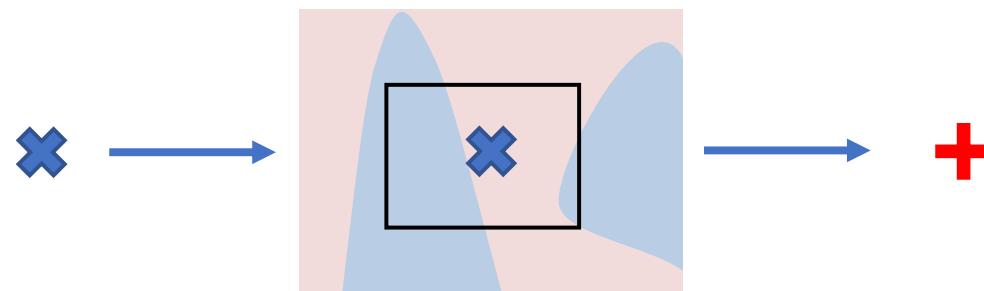


Explaining predictions



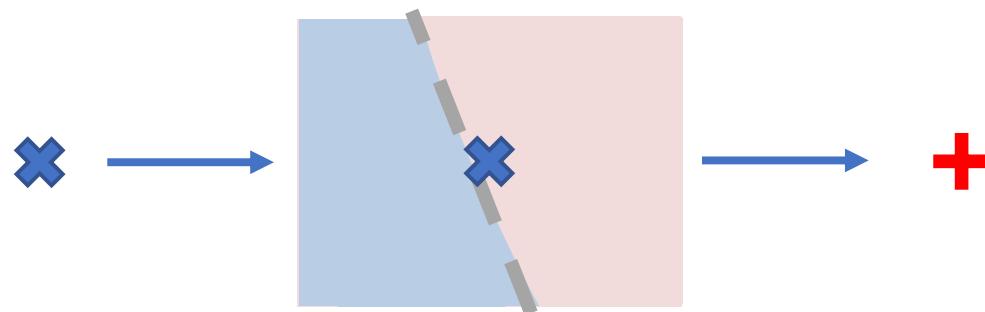
Explaining predictions

Locally, decision looks simpler...



Explaining predictions

Very locally, decision looks linear



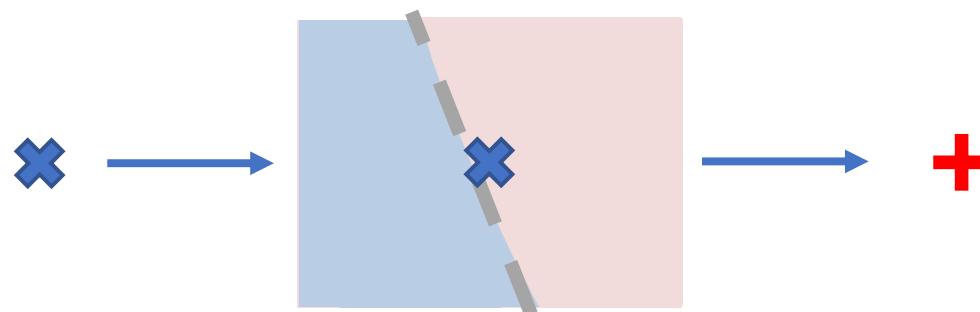
“Why should I trust you?”: Explaining the Predictions of Any Classifier. Ribeiro, Singh & G. KDD 16

CS229: Machine Learning

Explaining predictions

Very locally, decision looks linear

LIME: Learn locally sparse linear model around each prediction

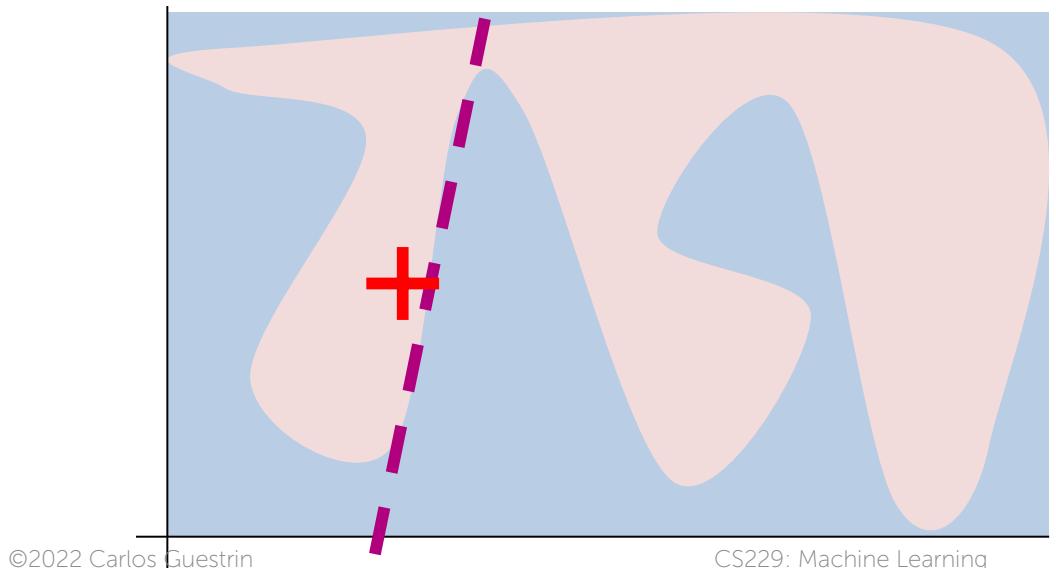


“Why should I trust you?”: Explaining the Predictions of Any Classifier. Ribeiro, Singh & G. KDD 16

CS229: Machine Learning

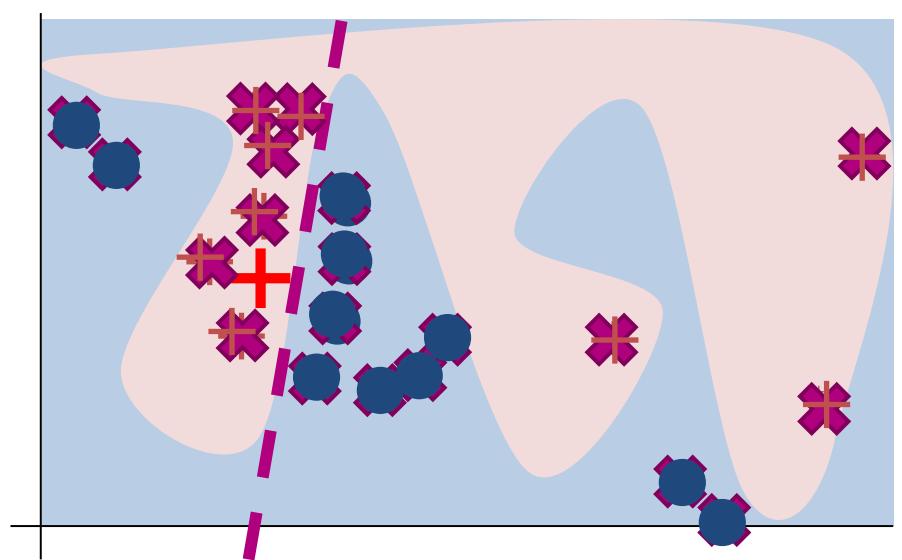
LIME – Key Ideas

1. Pick a model class interpretable by humans
2. Locally approximate global (blackbox) model
 - Simple model globally bad, but locally good



Sparse linear Explanations

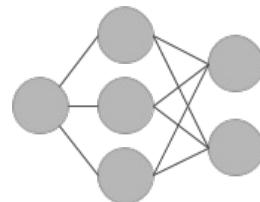
1. Sample points around x_i
2. Use complex model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn new simple model on weighted samples
5. Use simple model to explain



Interpretable representations

x (embeddings)

0.5	0.3	1.3	4.4	1.1	...
-----	-----	-----	-----	-----	-----



Model

x' (words)

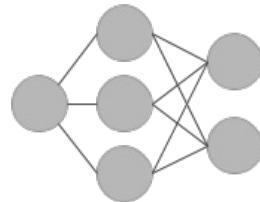
This is a horrible movie.



Human

Interpretable representation: images

x (3 color channels / pixel)



Model

x' (contiguous superpixels)



Human

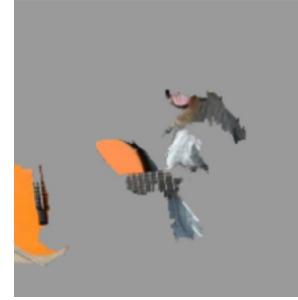
Explaining prediction of Inception Neural Network



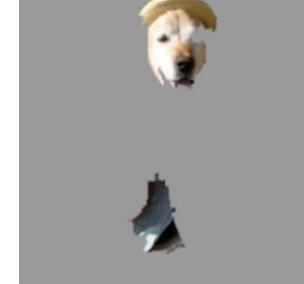
$P(\text{guitar}) = 0.32$



$P(\text{u}\text{guitar}) = 0.24$



$P(\text{dog}) = 0.21$



Achieving target metric may not be enough

Atheism vs Christianity posts
(Newsgroups data, circa 1995)



94% accuracy!!!

LIME applied to 20 newsgroups

From: Keith Jones
Subject: Christianity is the answer
NTP-Posting-Host: x.x.com

I think Christianity is the one true religion.
If you'd like to know more, send me a note

Appear in 21% of
training examples,
almost always in
Atheism

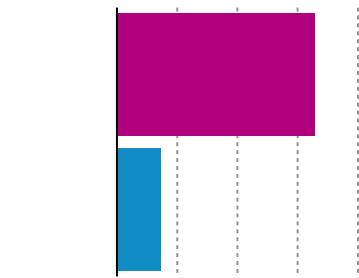
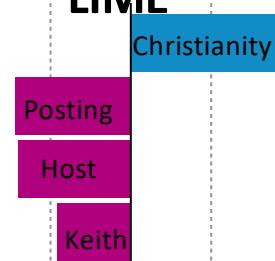
Appears in 11% of training
examples, **always** in atheism



Model



LIME



Prediction Prob.

Achieving target metric may not be enough

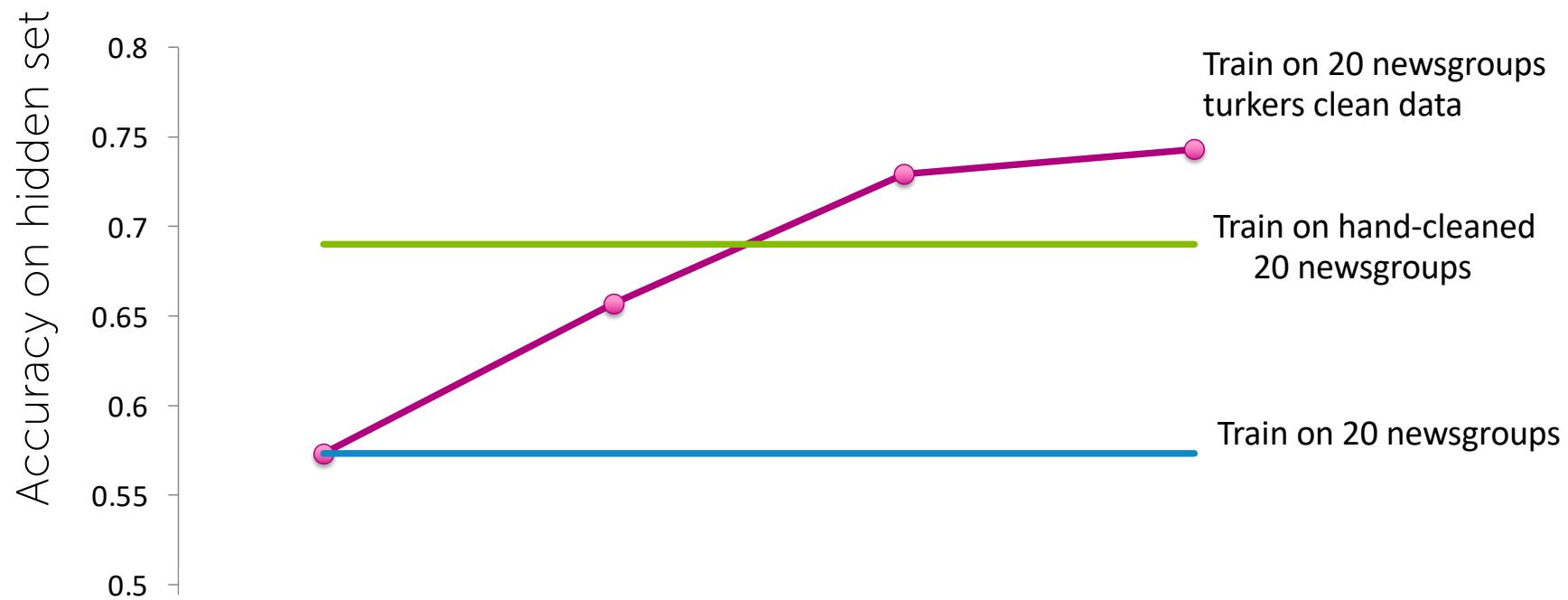
Atheism vs Christianity posts
(Newsgroups data, circa 1995)

94% accuracy!!!

Test on recent data:
Only 57% accuracy!

Predictions due to
email addresses, names,...

Fixing bad classifiers



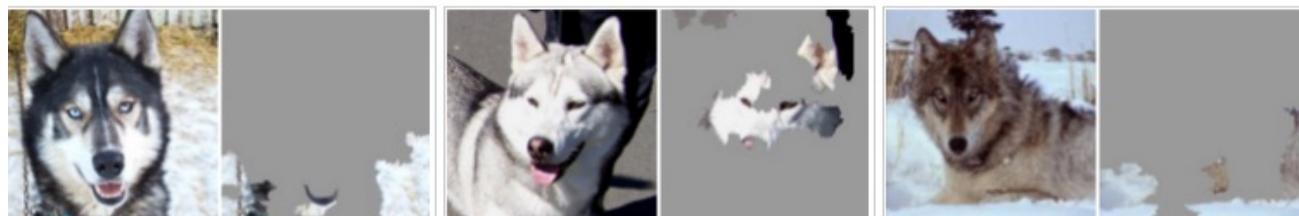


Predicted: **wolf**
True: **wolf**



Predicted: **husky**
True: **husky**

Predicted: **wolf**
True: **wolf**

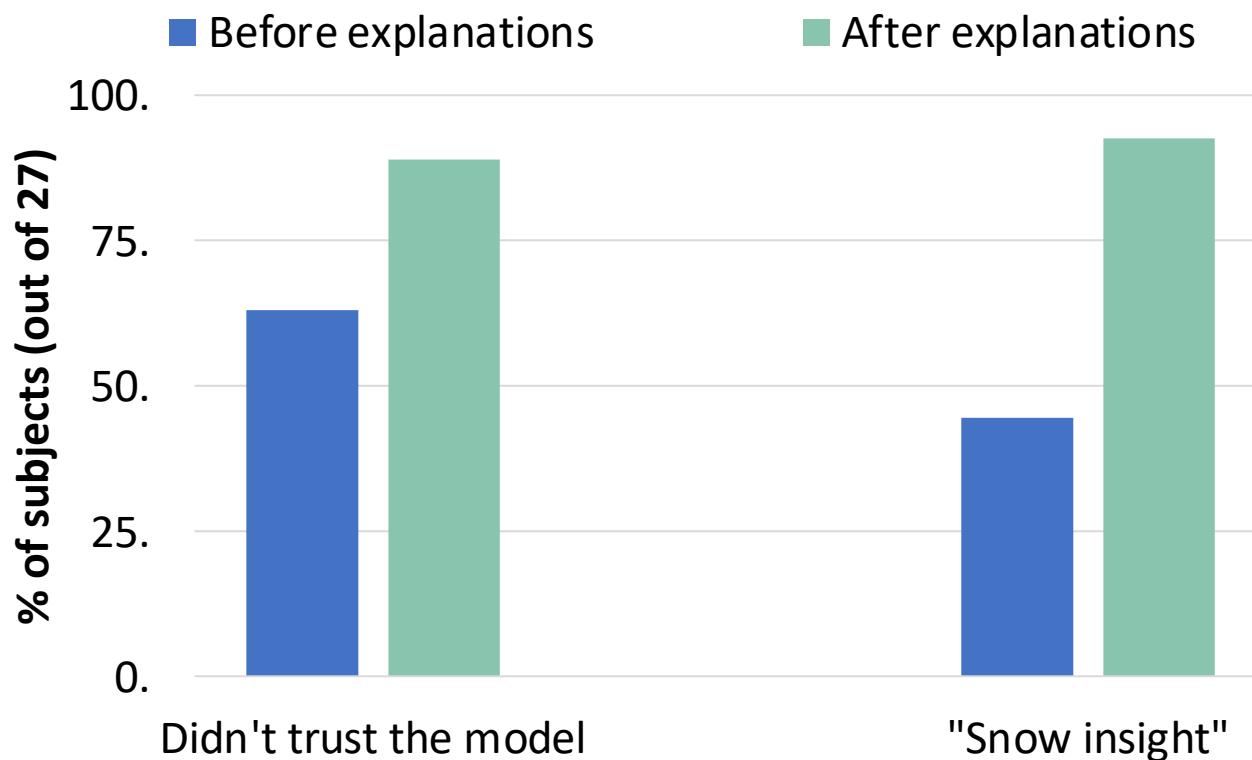


Predicted: **wolf**
True: **husky**

Predicted: **husky**
True: **husky**

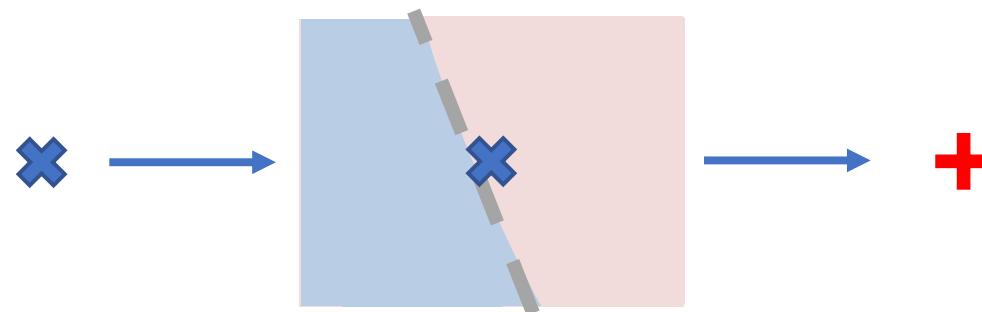
Predicted: **wolf**
True: **wolf**

Did explanations help with wolf problem?



More Examples

LIME: Learn locally sparse linear model around each prediction

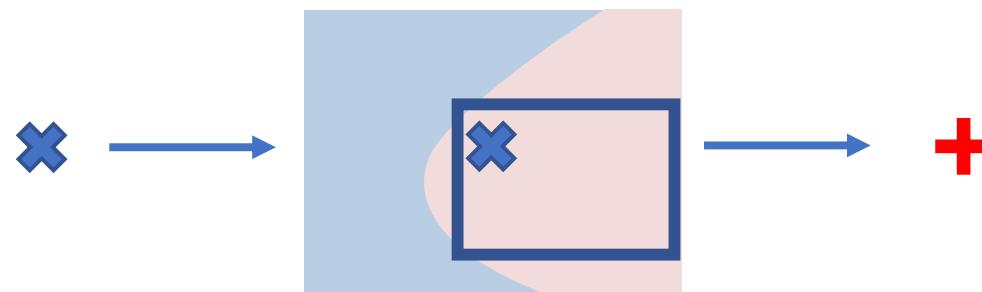


“Why should I trust you?”: Explaining the Predictions of Any Classifier. Ribeiro, Singh & G. KDD 16

CS229: Machine Learning

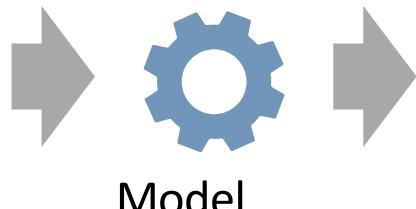
Anchors: Sufficient Conditions

Conditions under which classifier makes same prediction



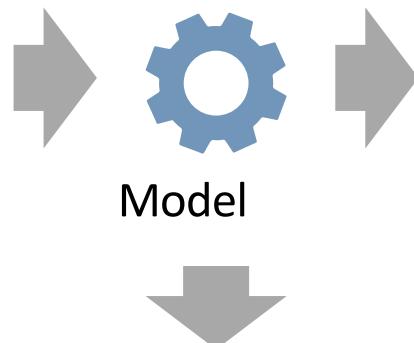
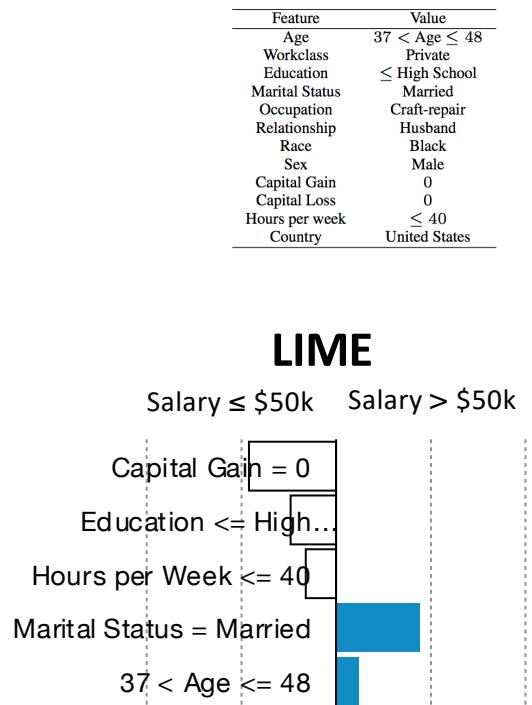
Salary Prediction

Feature	Value
Age	$37 < \text{Age} \leq 48$
Workclass	Private
Education	\leq High School
Marital Status	Married
Occupation	Craft-repair
Relationship	Husband
Race	Black
Sex	Male
Capital Gain	0
Capital Loss	0
Hours per week	≤ 40
Country	United States



Salary \leq \$50K

Salary Prediction: LIME vs Anchors



Salary \leq \$50K

Model

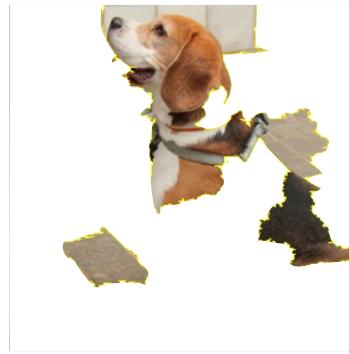
Anchor

IF Education \leq High School
Then
P(prediction = $\leq 50K$) $>$ 0.95

Anchors for Images: Classification



Prediction: Beagle



Anchor for Beagle

Anchors for Visual Question Answering



What is the mustache made of?

Banana

How many bananas are in the picture?

2

Anchors for Visual Question Answering

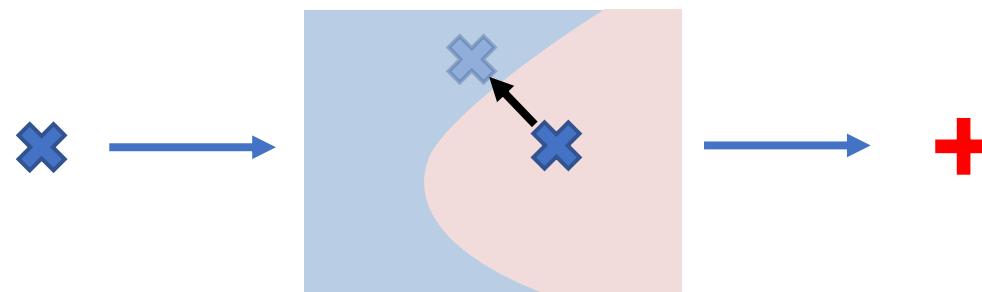


What is the mustache made of?	<i>Banana</i>
What is the ground made of?	<i>Banana</i>
What is the hair made of?	<i>Banana</i>
What is the picture of?	<i>Banana</i>
What was the head of the US?	<i>Banana</i>

How many bananas are in the picture?	2
How many are in the picture?	2
How many people in the picture?	2
Are there many animals in the picture?	2
How many is too many?	2

Adversarial Bug Discovery

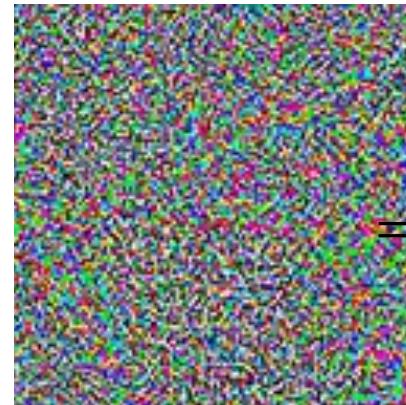
Find closest input with different prediction



Oversensitivity in image classification



+ ϵ



“Panda”

“Gibbon”

Adversary not distinguishable by human
→ Unlikely to be a real-world issue (except for attacks)



What type of road sign is shown?

STOP

The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe, at about 1,230 km.

How long is the Rhine?

1,230 km



The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe, at about 1,230 km.

What type of road sign is shown?
Which type of road sign is shown?

STOP

Do not enter

How long is the Rhine?
How long is the Rhine??

1,230 km

More than 1,050,000

Goal: Find semantically-equivalent adversarial examples

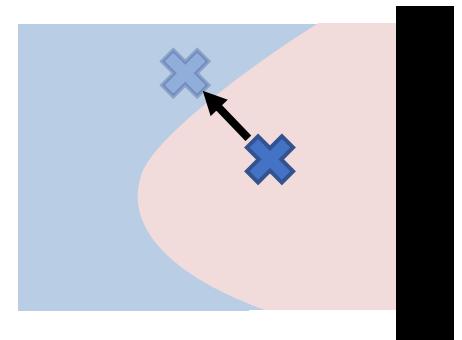
Semantically-equivalent
Use paraphrasing model
[Lapata et al. 2017]



Adversarial
Changes correct
model prediction



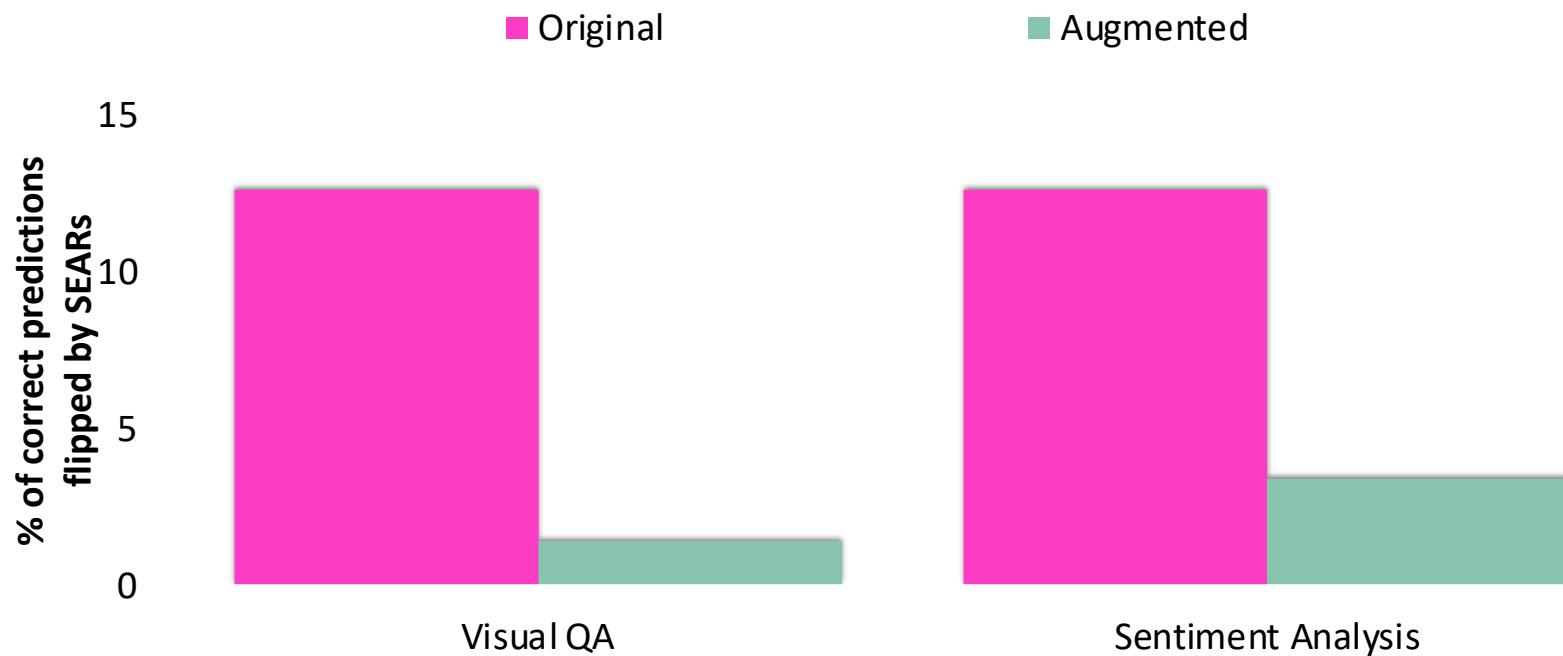
What color is the tray?	Pink
What colour is the tray?	Green
Which color is the tray?	Green
What color is it?	Green
What color is the tray?	Pink
How color is the tray?	Green



Closing the Loop with Simple Data

Augmentation

Augment by applying validated SEARs to training data



Typical challenges with explainability methods

- Explanations too simplistic
- Not focused on information needs for task
- Unstable
- Not causal
- ...

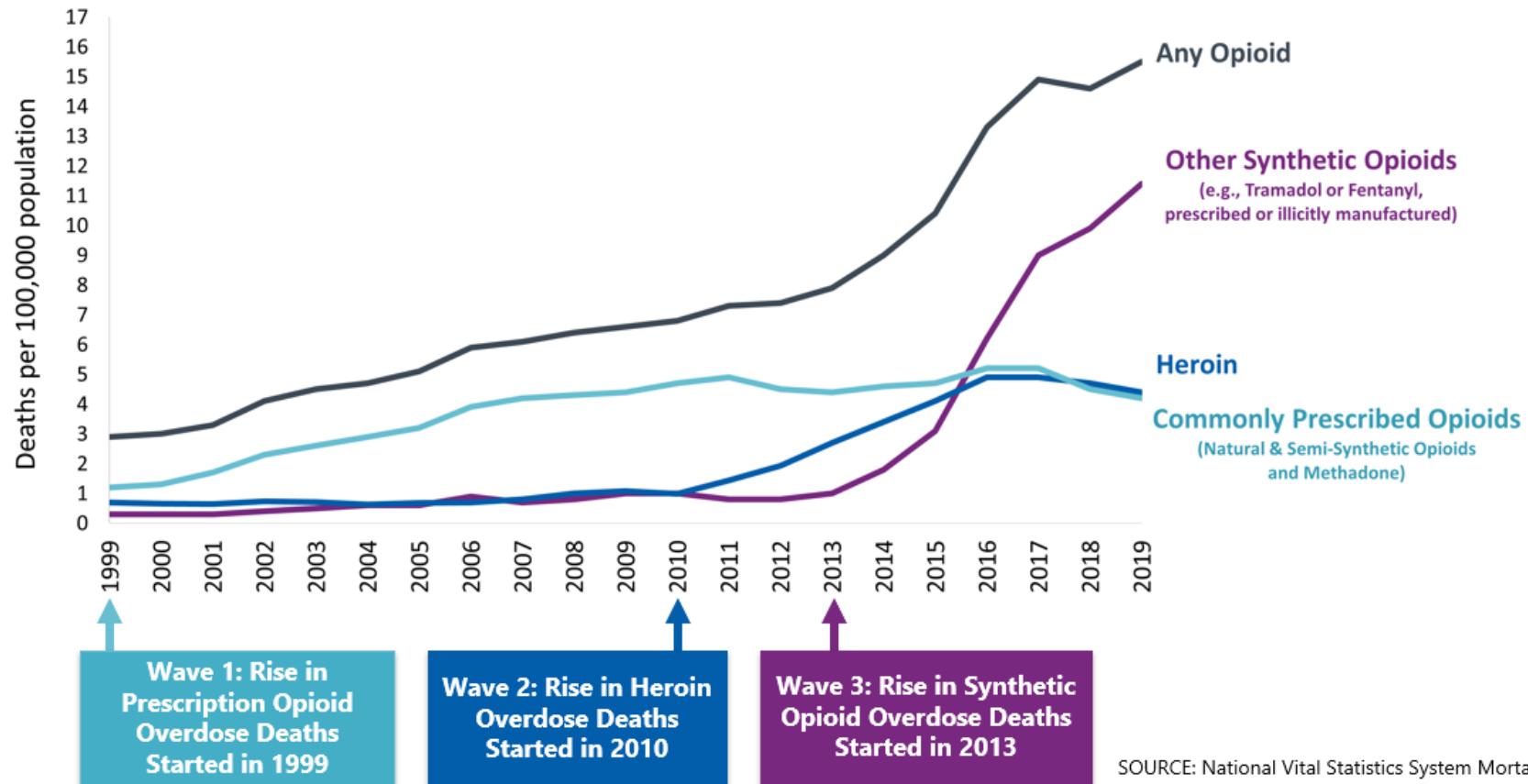
Annotated explainability Slides

AI Ethics: Explainability of Machine Learning

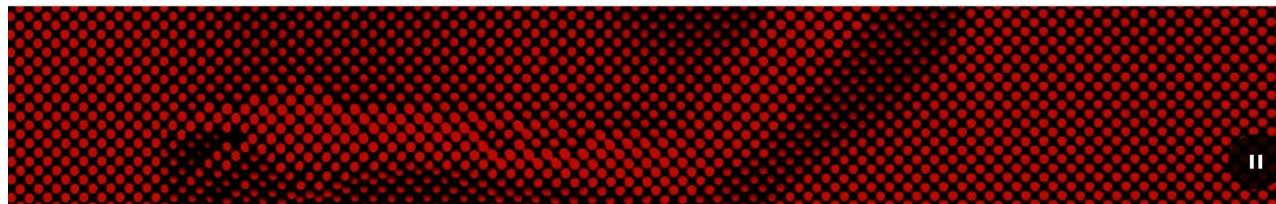
CS229: Machine Learning
Carlos Guestrin
Stanford University

©2022 Carlos Guestrin

Three Waves of the Rise in Opioid Overdose Deaths



SOURCE: National Vital Statistics System Mortality File.



VIDEO: SAM CANNON

MAIA SZALAVITZ

BACKCHANNEL AUG 11, 2021 6:00 AM

The Pain Was Unbearable. So Why Did Doctors Turn Her Away?

A sweeping drug addiction risk algorithm has become central to how the US handles the opioid crisis. It may only be making the crisis worse.

[The AI Database →](#)

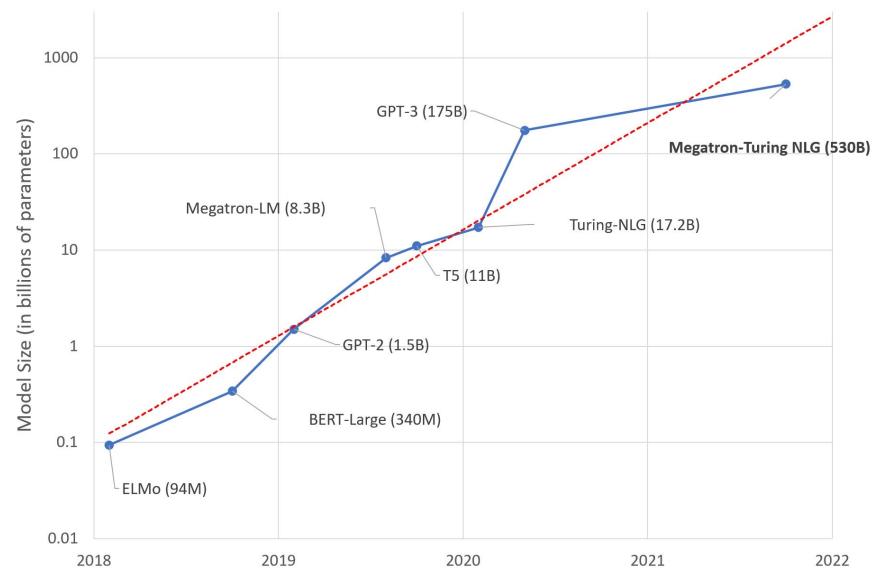
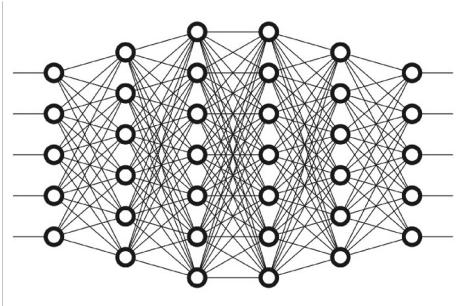
APPLICATION: ETHICS, PREDICTION, REGULATION

SECTOR: HEALTH CARE, PUBLIC SAFETY

ONE EVENING IN July of 2020, a woman named Kathryn went to the hospital in excruciating pain.

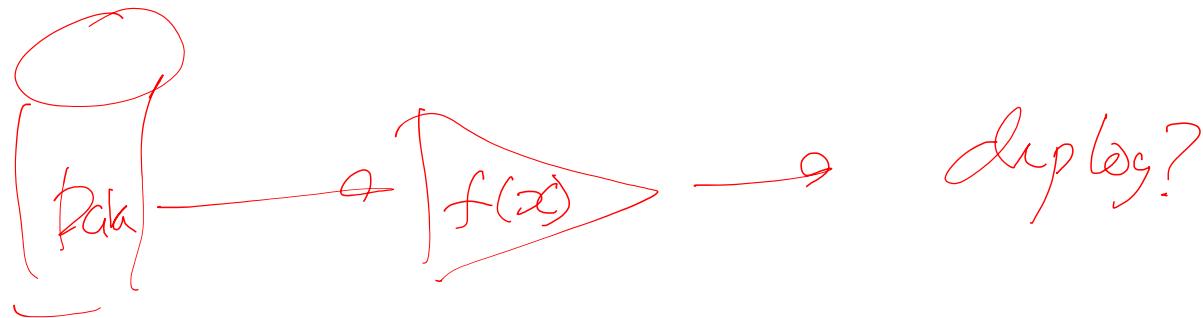
A 32-year-old psychology grad student in Michigan, Kathryn lived with endometriosis, an agonizing condition that causes uterine-like cells to abnormally develop in the wrong

ML Models More and More Complex



When is a model ready to deploy?

*Hard to understand when models are working
(for the right reasons) and not working!!*

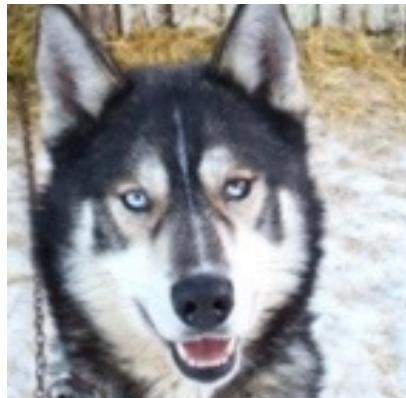


Isn't test accuracy enough?



A User Study on Test Accuracy

Train a neural network to predict **wolf** v. **husky**



Husky



Wolf

VIDEO

SLATE IN MOTION.

OCT. 14 2016 3:18 PM

The Man Who Accidentally Adopted a Wolf Pup

It did not go well.

By A.J. McCarthy



10k

547

6

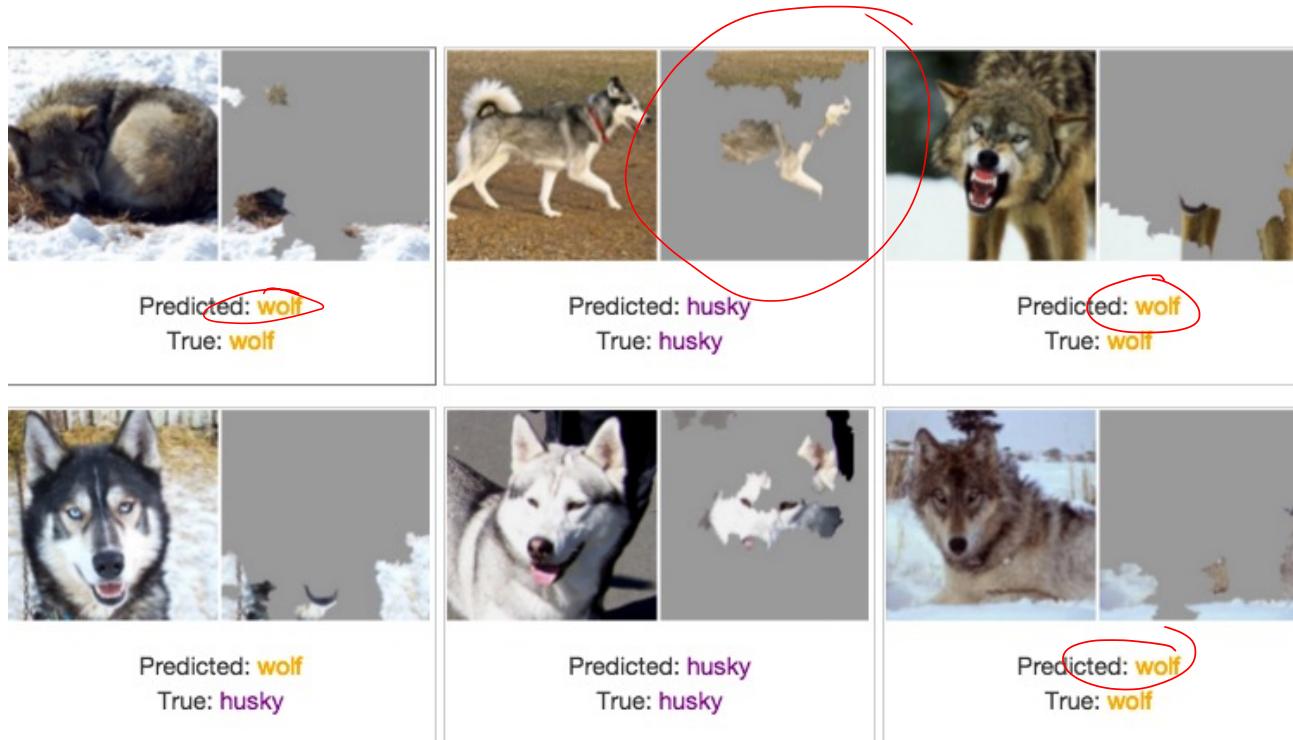


Train a neural network to predict wolf v. husky

					
Predicted: wolf True: wolf	Predicted: husky True: husky	Predicted: wolf True: wolf	Predicted: wolf True: husky	Predicted: husky True: husky	Predicted: wolf True: wolf

achieves desired accuracy

Explanations for neural network prediction



Snow detector

Test accuracy may not capture critical issues

- Bad data
- Biases
- Poor performance in critical cases
- ...

Examining Models

Debugging is One Reason to Examine Models

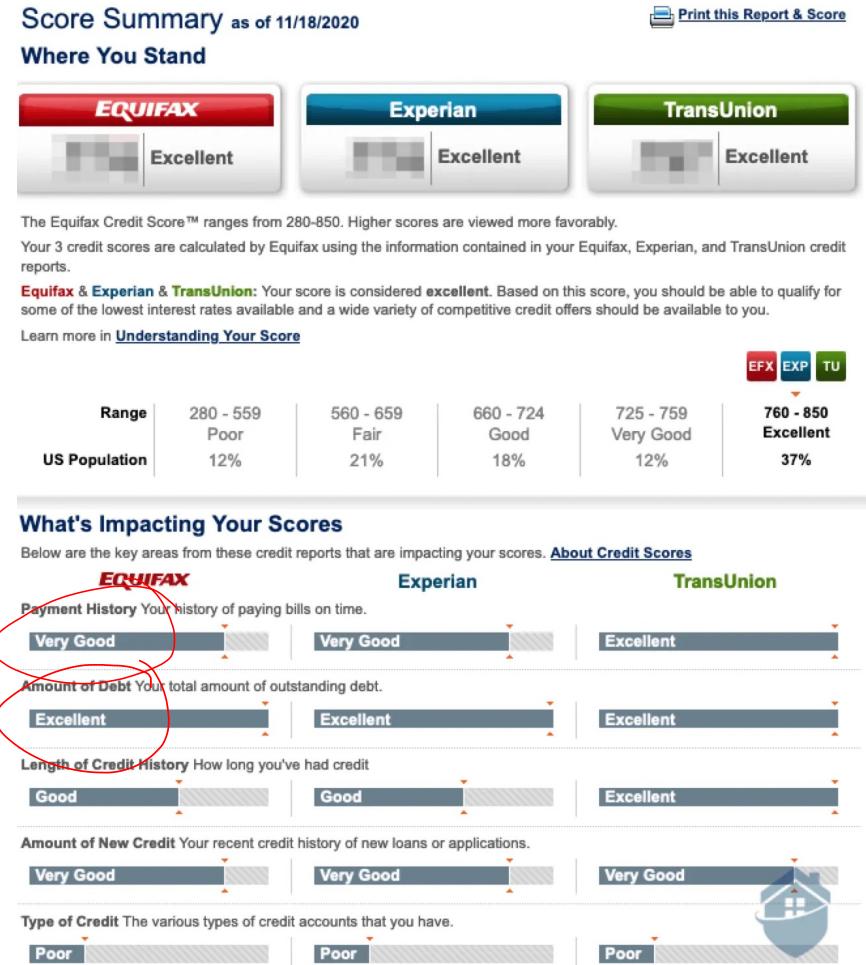
- Examining models:
 - Why a model makes particular predictions
 - What alternative predictions are possible
 - How robust/stable are predictions
 - What data supports predictions
- Examining models for debugging: discover bad, unexpected or unstable behavior
 - Typically not discovered by accuracy in train/test data

Examining Models to Detect Algorithmic Bias

- Evaluate multiple fairness criteria
- Verify how/if decisions depend on sensitive features
- Discover what groups are privileged/disadvantaged by predictions

Examine Models for Recourse

- In opioid overdose risk case, patient deemed risky had no way to discover why
 - Or how to fix bad data
- Understanding why could enable individuals to:
 - Address data issues
 - Change their actions to change outcomes



Interpretable Models vs Post-hoc Explanations

Interpretability in ML

*Giving humans a **mental model** of
the machine's model behavior*

Learning Interpretable Models

(c.f., Lethan & Rudin 2015)

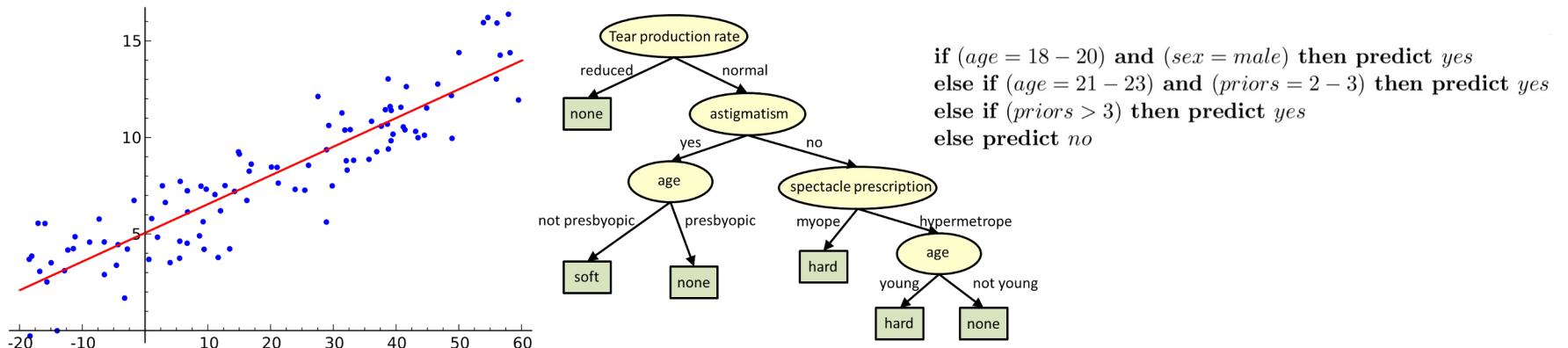
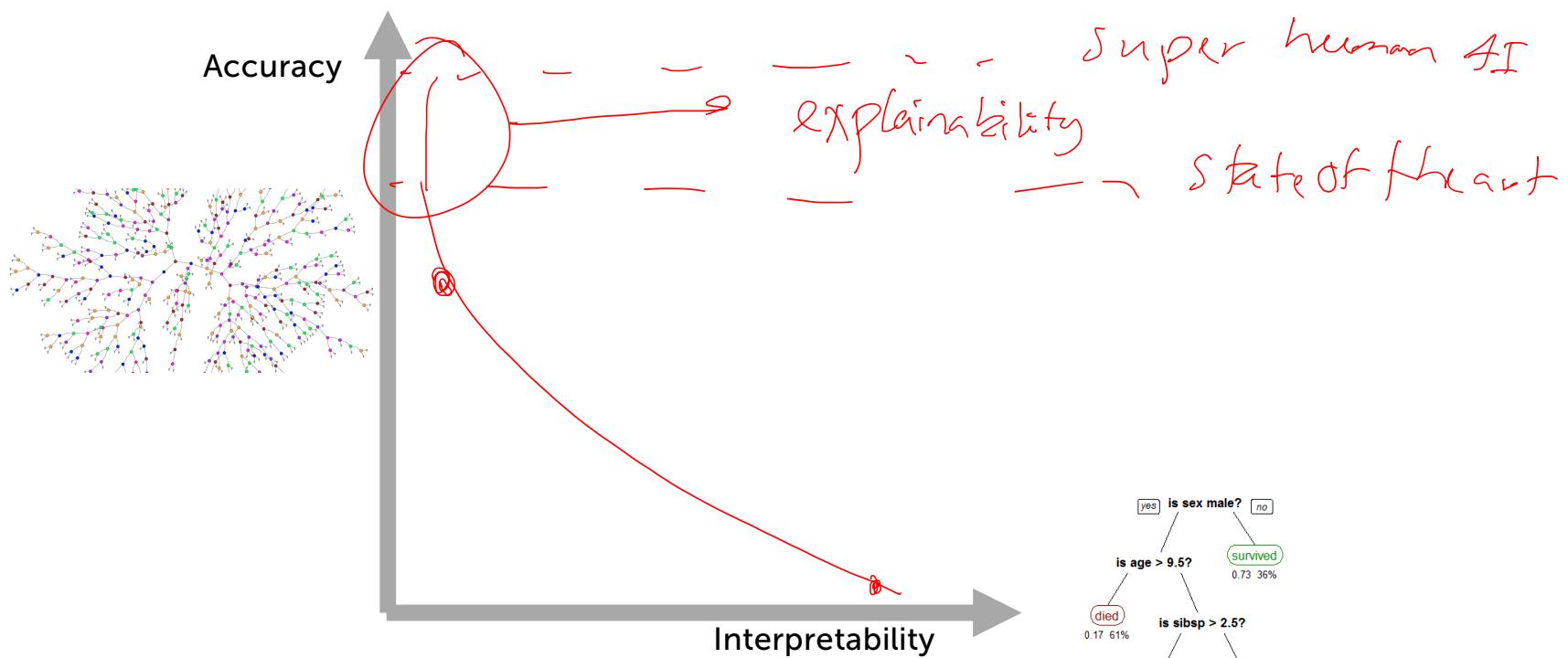


Image credit: Lakkaraju, Adebayo, Singh NeurIPS 2020 Tutorial

Accuracy vs Interpretability



©2022 Carlos Guestrin

CS229: Machine Learning

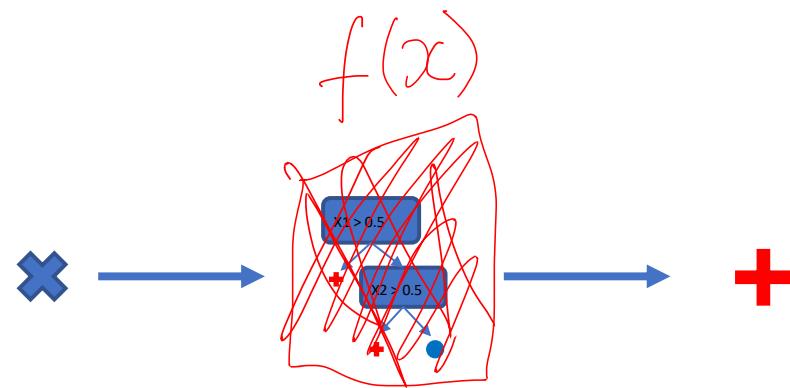
Post-hoc Explanations

- Given a (huge, complex) model, provide human explanations for predictions



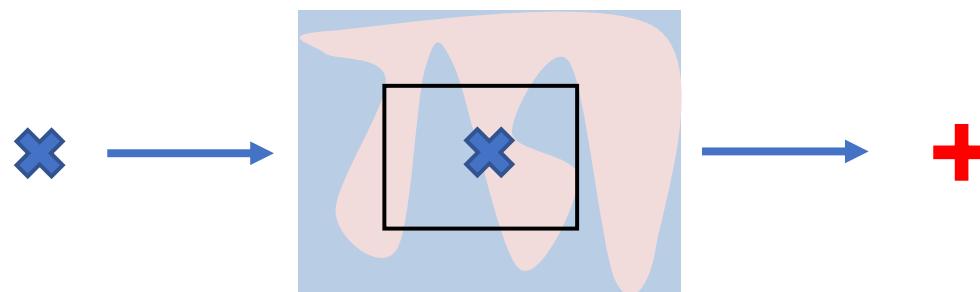
LIME: Local, Interpretable Model-Agnostic Explanations

Model agnostic → Ignore any internal structure



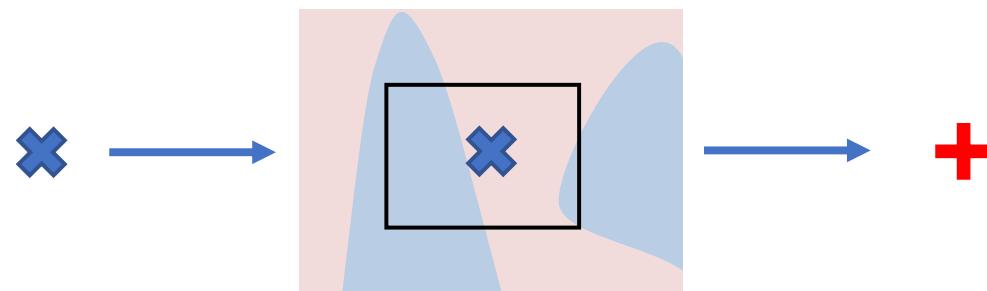
Compare different function class
state-of-the-art
future proof

Explaining predictions



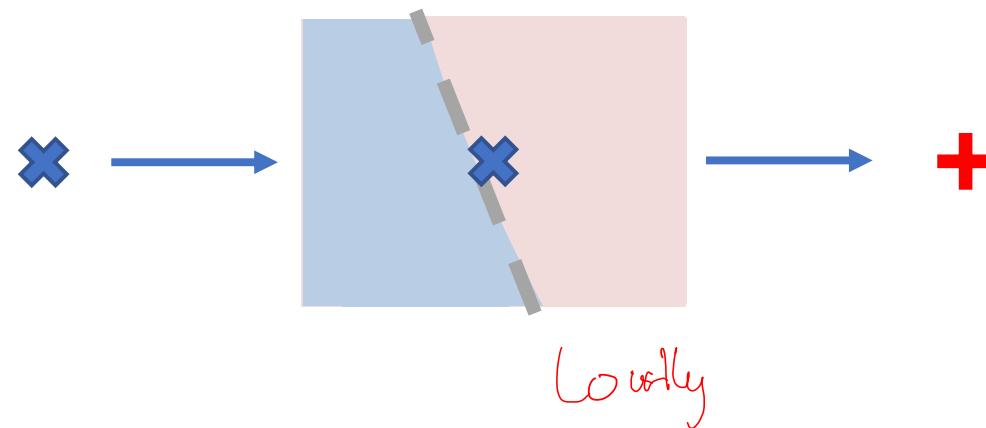
Explaining predictions

Locally, decision looks simpler...



Explaining predictions

Very locally, decision looks linear



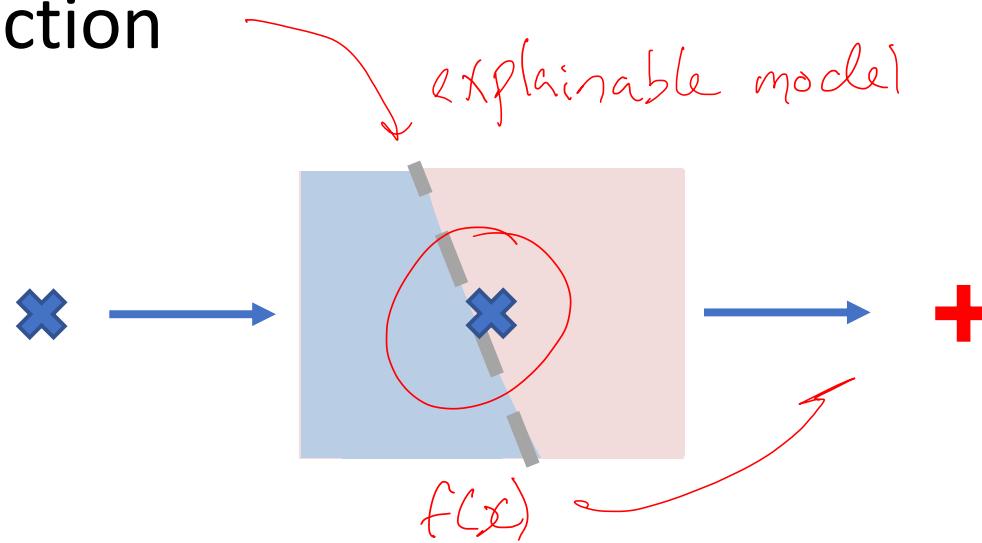
“Why should I trust you?”: Explaining the Predictions of Any Classifier. Ribeiro, Singh & G. KDD 16

CS229: Machine Learning

Explaining predictions

Very locally, decision looks linear

LIME: Learn locally sparse linear model around each prediction

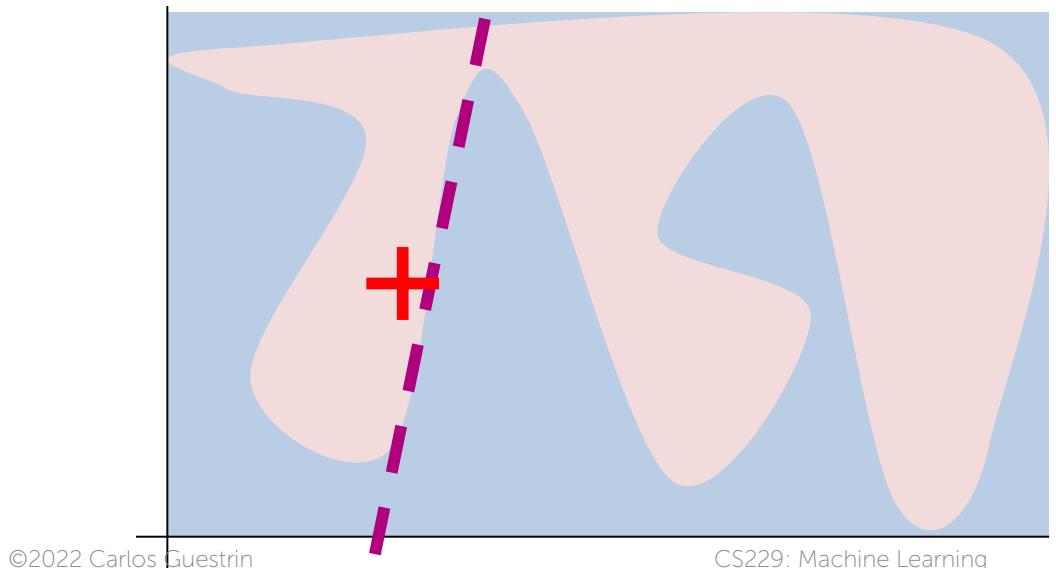


“Why should I trust you?”: Explaining the Predictions of Any Classifier. Ribeiro, Singh & G. KDD 16

CS229: Machine Learning

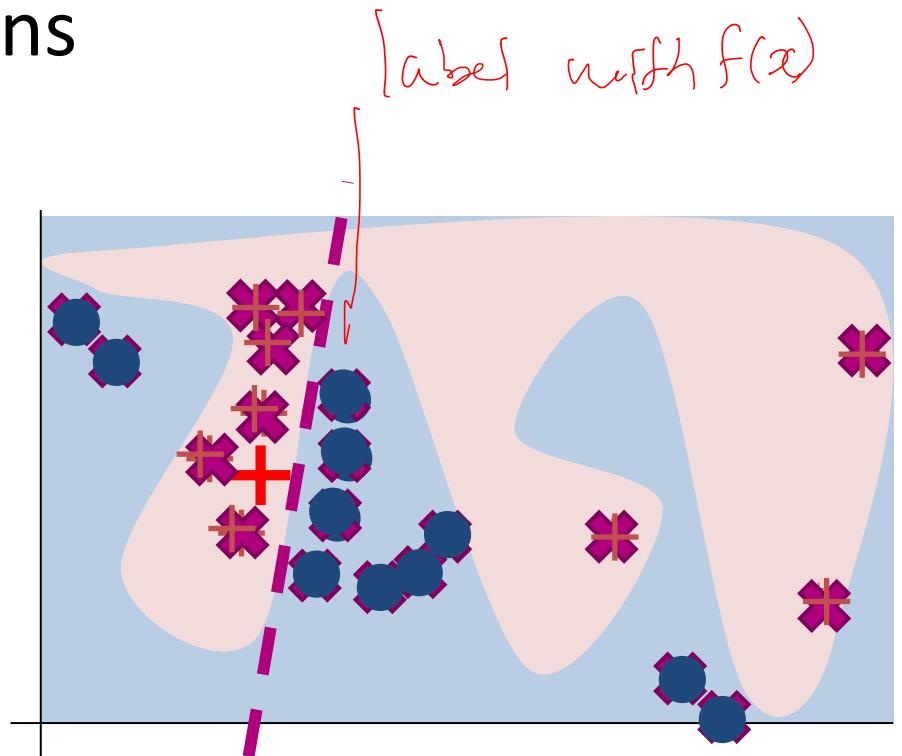
LIME – Key Ideas

1. Pick a model class interpretable by humans
2. Locally approximate global (blackbox) model
 - Simple model globally bad, but locally good



Sparse linear Explanations

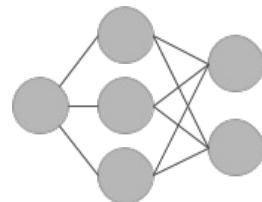
1. Sample points around x_i
2. Use complex model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn new simple model on weighted samples
5. Use simple model to explain



Interpretable representations

x (embeddings)

0.5	0.3	1.3	4.4	1.1	...
-----	-----	-----	-----	-----	-----



Model

x' (words)

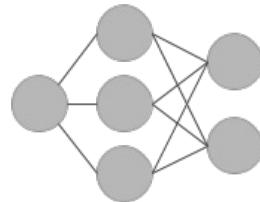
This is a horrible movie.



Human

Interpretable representation: images

x (3 color channels / pixel)



Model

x' (contiguous superpixels)



Human

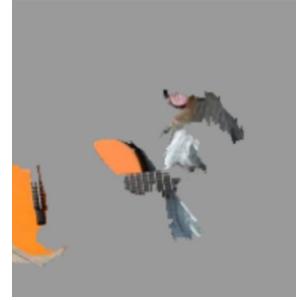
Explaining prediction of Inception Neural Network



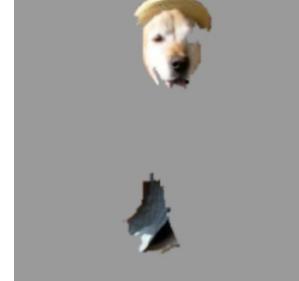
$$P(\text{Guitar}) = 0.32$$



$$P(\text{Guitar}) = 0.24$$



$$P(\text{Dog}) = 0.21$$



Achieving target metric may not be enough

Atheism vs Christianity posts
(Newsgroups data, circa 1995)



94% accuracy!!!

LIME applied to 20 newsgroups

From: Keith Jones
Subject: Christianity is the answer
NTP-Posting-Host: x.x.com

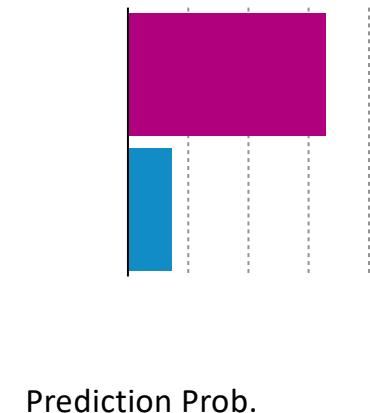
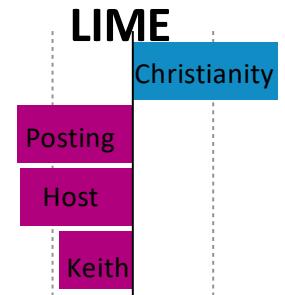
I think Christianity is the one true religion.
If you'd like to know more, send me a note

Appear in 21% of
training examples,
almost always in
Atheism

Appears in 11% of training
examples, **always** in atheism



Model



Achieving target metric may not be enough

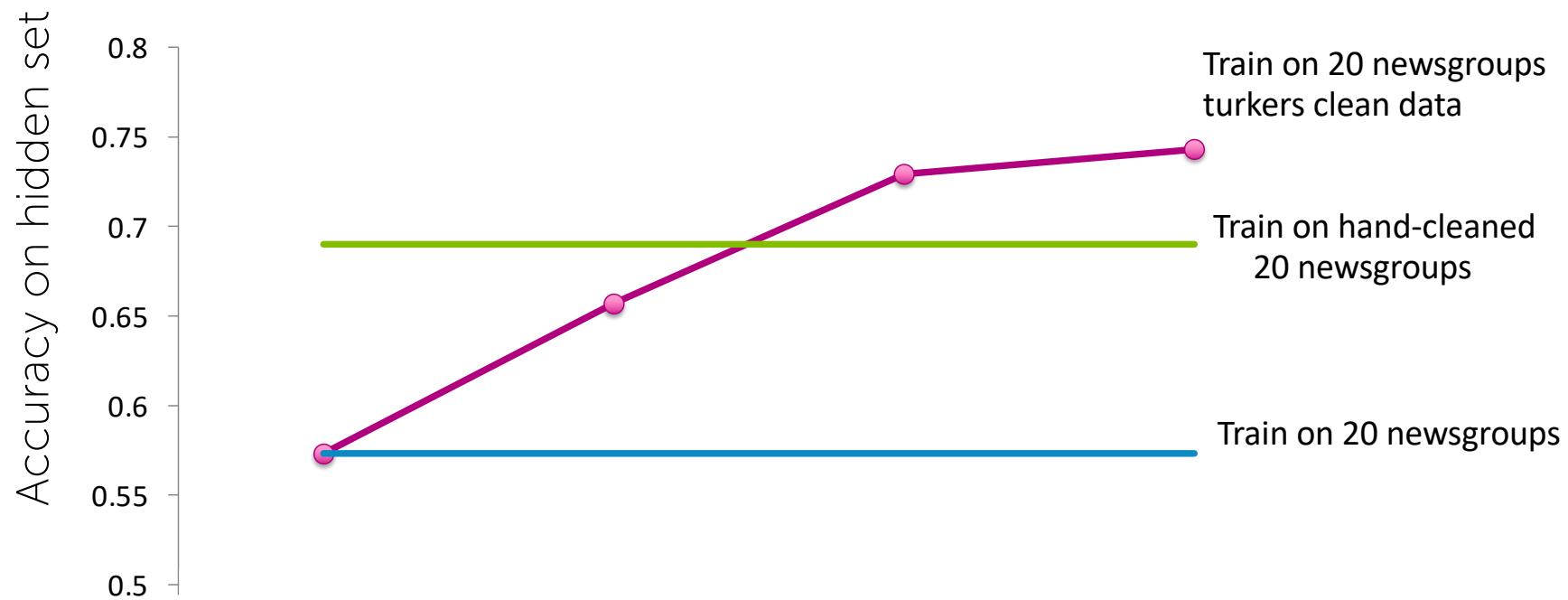
Atheism vs Christianity posts
(Newsgroups data, circa 1995)

94% accuracy!!!

Test on recent data:
Only 57% accuracy!

Predictions due to
email addresses, names,...

Fixing bad classifiers



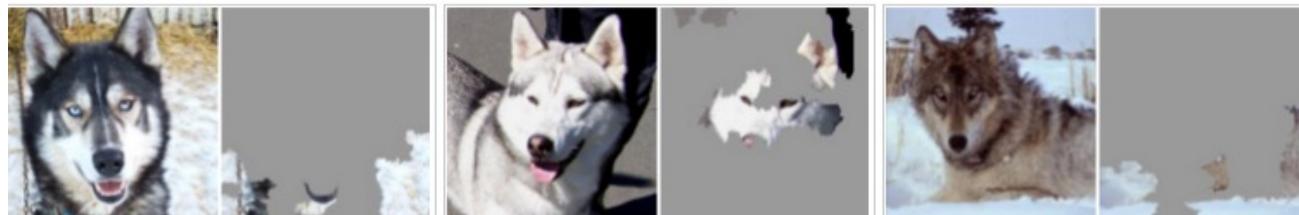


Predicted: **wolf**
True: **wolf**



Predicted: **husky**
True: **husky**

Predicted: **wolf**
True: **wolf**

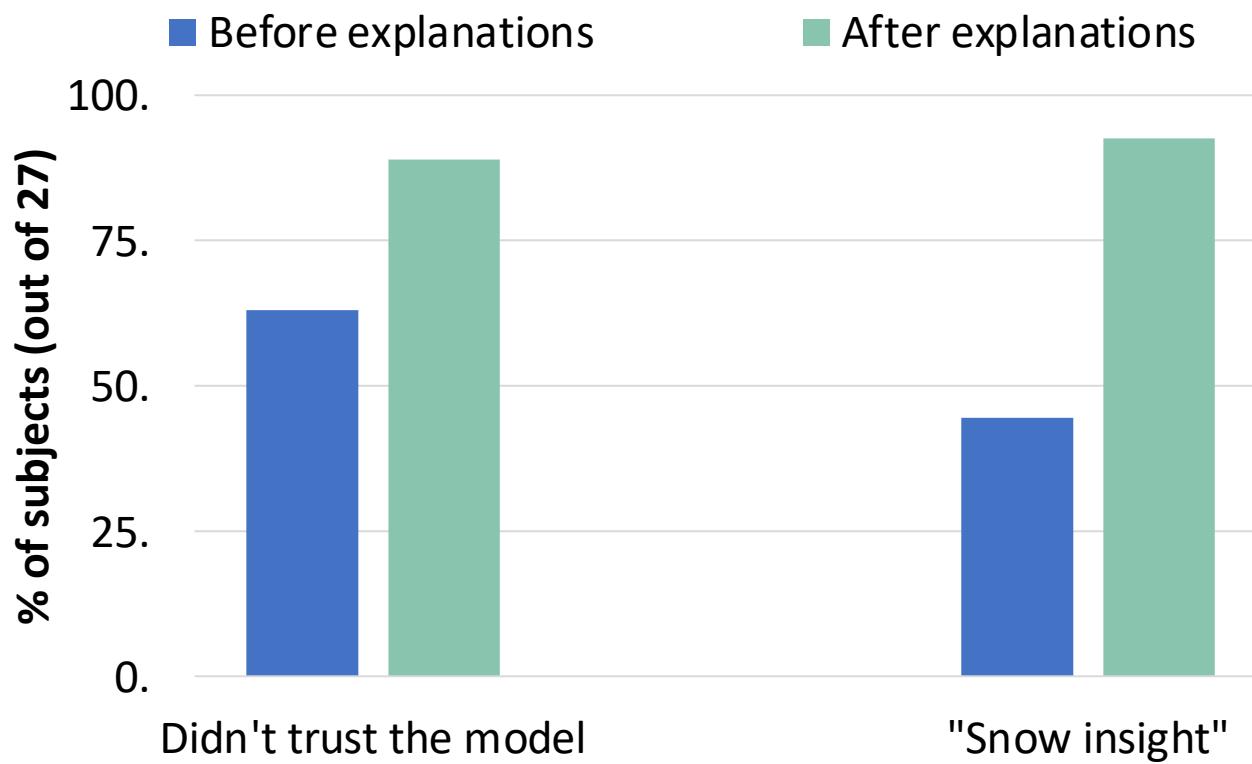


Predicted: **wolf**
True: **husky**

Predicted: **husky**
True: **husky**

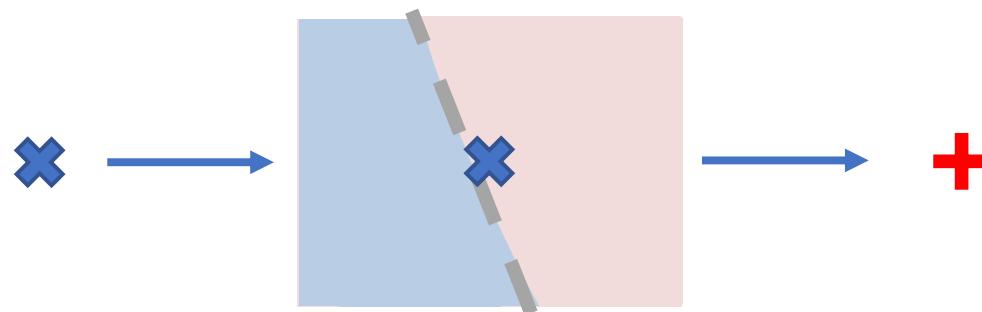
Predicted: **wolf**
True: **wolf**

Did explanations help with wolf problem?



More Examples

LIME: Learn locally sparse linear model around each prediction

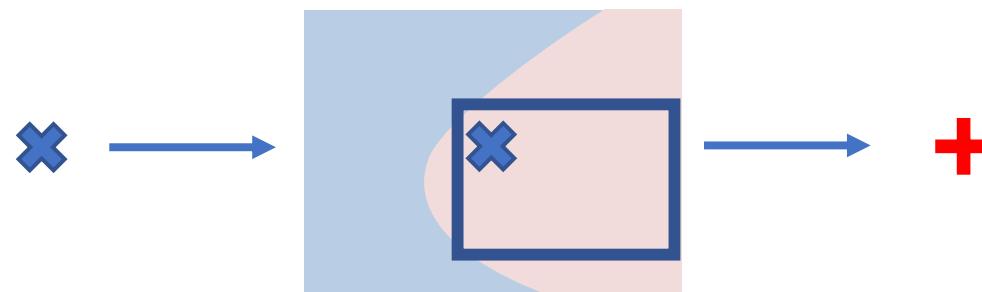


“Why should I trust you?”: Explaining the Predictions of Any Classifier. Ribeiro, Singh & G. KDD 16

CS229: Machine Learning

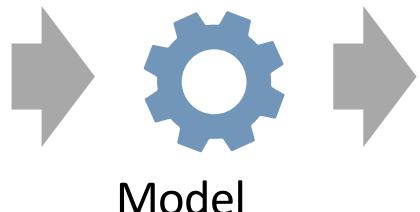
Anchors: Sufficient Conditions

Conditions under which classifier makes same prediction



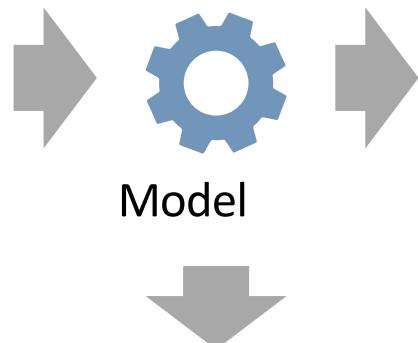
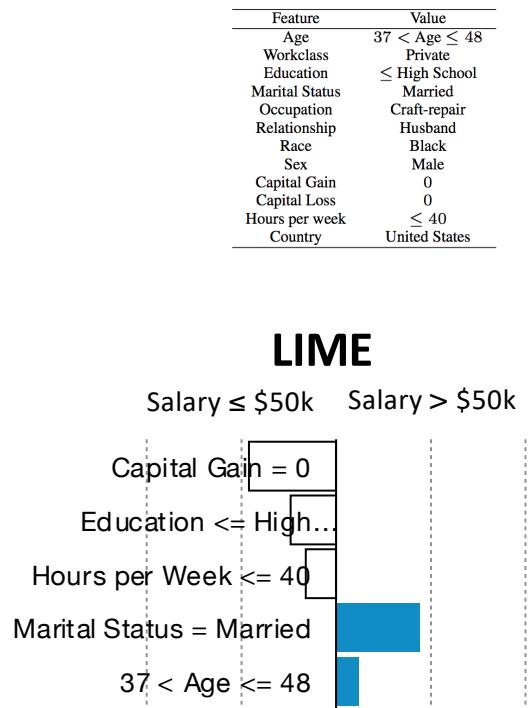
Salary Prediction

Feature	Value
Age	$37 < \text{Age} \leq 48$
Workclass	Private
Education	\leq High School
Marital Status	Married
Occupation	Craft-repair
Relationship	Husband
Race	Black
Sex	Male
Capital Gain	0
Capital Loss	0
Hours per week	≤ 40
Country	United States



Salary \leq \$50K

Salary Prediction: LIME vs Anchors



Salary \leq \$50K

Model

Anchor

IF Education \leq High School
Then
P(prediction = $\leq 50K$) $>$ 0.95

Anchors for Images: Classification



Prediction: Beagle



Anchor for Beagle

Anchors for Visual Question Answering



What is the mustache made of?	Banana
-------------------------------	--------

Anchor

How many bananas are in the picture?	2
--------------------------------------	---

Anchors for Visual Question Answering

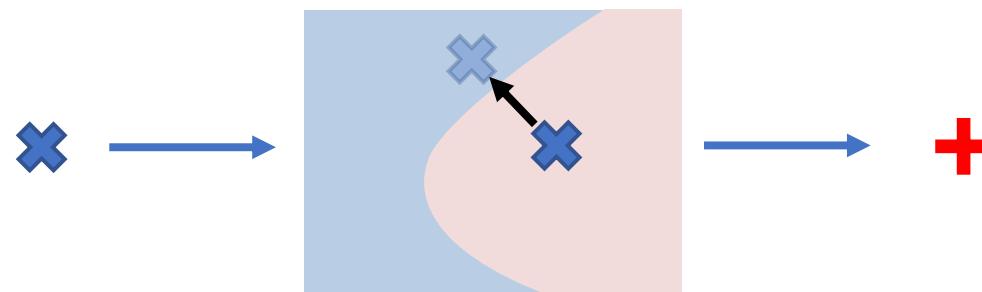


What is the mustache made of?	<i>Banana</i>
What is the ground made of?	<i>Banana</i>
What is the hair made of?	<i>Banana</i>
What is the picture of?	<i>Banana</i>
What was the head of the US?	<i>Banana</i>

How many bananas are in the picture?	2
How many are in the picture?	2
How many people in the picture?	2
Are there many animals in the picture?	2
How many is too many?	2

Adversarial Bug Discovery

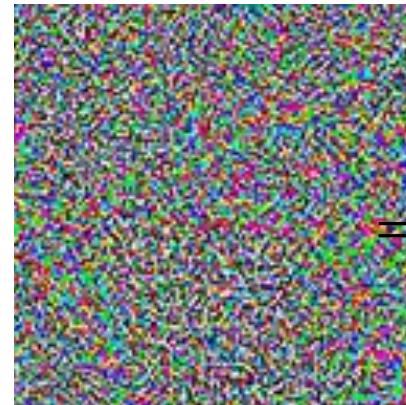
Find closest input with different prediction



Oversensitivity in image classification



+ ϵ



“Panda”

“Gibbon”

Adversary not distinguishable by human
→ Unlikely to be a real-world issue (except for attacks)



What type of road sign is shown?

STOP

The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe, at about 1,230 km.

How long is the Rhine?

1,230 km



The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe, at about 1,230 km.

What type of road sign is shown?
Which type of road sign is shown?

STOP

Do not enter

How long is the Rhine?
How long is the Rhine??

1,230 km

More than 1,050,000

Goal: Find semantically-equivalent adversarial examples

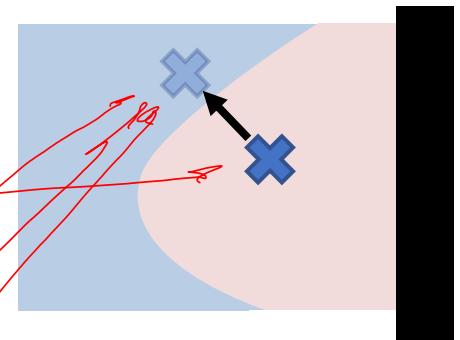
Semantically-equivalent
Use paraphrasing model
[Lapata et al. 2017]



Adversarial
Changes correct
model prediction



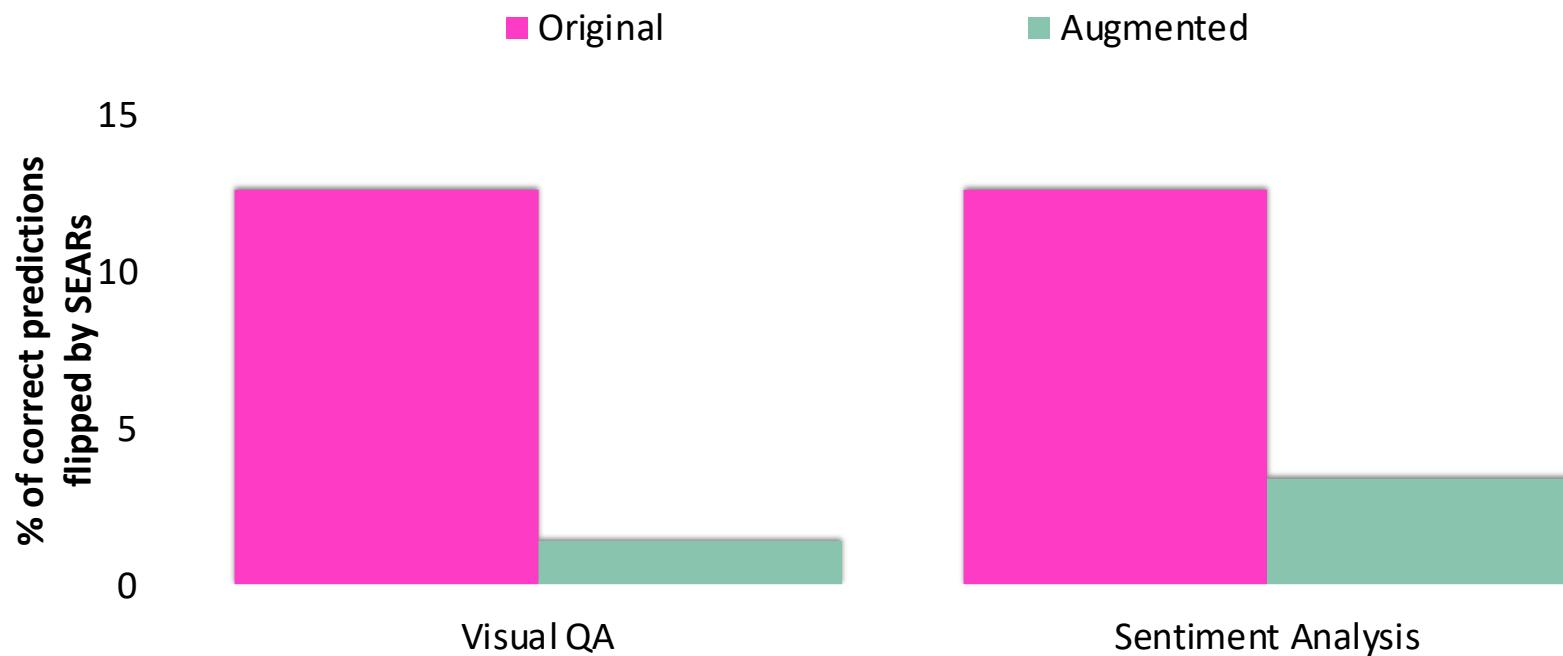
What color is the tray?	Pink
What colour is the tray?	Green
Which color is the tray?	Green
What color is it?	Green
What color is the tray?	Pink
How color is the tray?	Green



Closing the Loop with Simple Data

Augmentation

Augment by applying validated SEARs to training data



Typical challenges with explainability methods

- Explanations too simplistic
- Not focused on information needs for task
- Unstable
- Not causal
- ...